Building a comparable corpus of online discussions on Wikipedia

Lydia-Mai Ho-Dac

Building a comparable corpus of online discussions on Wikipedia: The EFG WikiCorpus

Lydia-Mai Ho-Dac

Building corpora from Wikipedia that are reusable and can be employed for linguistic analyses in the same way as traditional corpora is an exciting challenge, especially if we consider not only the articles but also the discussions that take place behind the scenes of the collaborative encyclopedia. Wikipedia is self-organized into different "namespaces" that correspond to different genres: encyclopedic articles, thematic portal descriptions, help pages and different kinds of discussions ("talk pages" in Wikipedia terms); these include talks about cowriting an article, talks focusing on a Wikipedia user, and talks for welcoming new users or discussing Wikipedia policies.

Wikipedia (henceforth WP) is an overwhelming place where human knowledge is recorded, shared, discussed and, in a certain way, built. The well-known DBpedia project takes full advantage of such a resource by aiming to build a kind of world ontology of human knowledge (Lehman et al. 2012). WP is an artifact that exists only in the collaborative process flow that gave rise to it. Unlike other published encyclopedias, there are no (and will never be) stable releases of WP. In other words, WP pages will never reach "final version" status because WP content follows a very different editorial process from the usual one: submission, revision, rewriting and publication.

The evolution of the encyclopedia is and will always be accessible via the WP revision history of all pages, including articles and discussions. WP's revision history is one of its key features. It corresponds to a complete record of all editing actions, i.e., the addition and deletion of content - carried out by any user on any page since WP's inception. The existence of the WP revision history allows any user to "revert", i.e., to restore one part of the page to a previous version. As a consequence, WP is an exciting resource for studying collaborative writing processes across digital humanities. On the one hand, the WP revision history provides new insights into writing and revision processes (cf. Ferschke et al. 2013, Borra et al. 2015). It also provides data for improving vandalism and harassment detection (cf. Potthast et al. 2008, Wulczyn et al. 2017). On the other hand, the talk pages, i.e., the discussions that take place behind the articles, offer scientists an unexpected chance to observe how users interact and what topics they explicitly discuss during the collaborative writing process (cf. Ferschke et al. 2013).

This chapter presents a compilation process that was used to build the EFG WikiCorpus, a comparable corpus composed of all the talk pages in the English, French and German WP. Only talk pages dedicated to cowriting articles were selected. The resulting EFG WikiCorpus contains more than 3 million talk pages with more than 2 billion words. The compilation process starts from the WP archives that are regularly provided as database dumps on the web by the Wikimedia foundation. These source data, which were obtained in a specific format, the wikicode, have been parsed and converted according to the TEI CMC-core schema, using ancillary techniques for structuring the content and extracting the metadata of a talk page. This chapter describes the outlines of the building process and presents

statistics of the EFG WikiCorpus and the EFG WikiDemoCorpus (WDC), a derived subcorpus used for qualitative analyses in different contributions of this volume.

This chapter describes the structure of the talk pages across the three languages (Section 2) and explains how we compiled and structured the EFG WikiCorpus corpus and encoded it according to the TEI CMC-core schema (Section 3). The chapter ends by providing descriptive statistics of the resulting corpora and a recommendation for selecting subcorpora such as the EFG WikiDemoCorpus as used in different contributions of this volume.

1- Wikipedia talk pages: Wikipedia's backstage

WP talk pages, or discussions, are considered WP's backstage. Such discussions are crucial for ensuring the self-managing organization that is typical of WP and that Konieczny (2010) defines as an "adhocracy" on the basis of Mintzberg's models (1979) of adhocratic governance. For Konieczny (2010:10), WP is a successful self-evolved organizational structure that "works in reality, not in theory":

"As there are no official "Wikipedia employees," the site's entire governance structure, managing millions of volunteers working on a similar number of content pages, has been created by its on-line volunteers. Wikipedia allows all its editors to vote and voice their opinions, and empowers them to change the content of articles and of organizational policies to an extent unthinkable in traditional organizations."

Talk pages are a cornerstone of such a "Wikicracy", supported as the "democracy of the future" by some Wikipedians (cf. Wikimedia (2009). Wikicracy. Retrieved on 4 March 2019[1]). It is where Wikpedians "vote and voice their opinions, and [this] empowers them to change the content of articles".

For instance, on the talk page of the English article about "Corpus Linguistics" named "Talk:Corpus Linguistics"[2], the user @Hutschi opened a discussion in 2004 headed "Chomsky and Corpus Linguistics" by asking[3]:

(1)

> "Chomsky and Corpus Linguistics
> The article states: The approach runs counter to Noam Chomsky's view that real language is riddled with performance-related errors, thus requiring careful analysis of small speech samples obtained in a highly controlled laboratory setting. When did Chomsky say this and where? Do the two appoaches contradict each other or do they complete each other? --Hutschi 10:47, 7 Jul 2004 (UTC)"

This question is followed by several asynchronous answers and explanations, the first one in April 2007 and the last one in April 2010. As this example shows, talk pages constitute a sort of online forum on which Wikipedians discuss the ongoing writing process with other Wikipedians. As emphasized by Ferschke (2014:11), a WP talk page corpus is an unparalleled observatory of human collaboration:

[1] https://meta.wikimedia.org/wiki/Wikicracy
[2] https://en.wikipedia.org/wiki/Talk:Corpus_linguistics
[3] The misspelling "appoaches" is as in the original.

"From a scientific point of view, article Talk pages are a unique type of web discourse and a valuable resource for the humanities and writing sciences, since the discussions develop in parallel with the discussed articles and provide insights into the meta level of the collaborative writing process that normally remains hidden. With structured access to this resource, the linguists and researchers in the writing sciences have the unparalleled possibility to observe these hidden processes without having to conduct interviews or carrying out supervised field experiments."

Talk pages were not originally included in Wikipedia. A few days after the launch of the English WP (WP.en), a Wikipedian raised the following question: "What to do with discussions behind the articles?". The first reaction of @Jimbo was to answer that WP as a topic must be discussed in another place (e.g., on mailing lists), as stated in the Jimbo Wales Statement of principles[4]:

"Wikipedia is an encyclopedia. The topic of Wikipedia articles should always look outward, not inward at Wikipedia itself."

Nevertheless, a place dedicated to discussion and negotiation rapidly became crucial, especially for developing non-English WPs that are out of Wales' control. The first non-English WPs were the German and French language versions (WP.de and WP.fr) that launched in March 2001 and the Italian WP (WP.it) that launched in May 2001. Almost immediately after the WP.de kickoff, WP.de users developed forums, e.g., in German named *Meinungsbilder* ("Meinung+bilder" for "opinion+builder") for clarifying issues for which there is no consensus. In the same vein, Florence Dévouard, alias @Anthère on WP.fr, created a page called *Decision-making* in October 2002 where "the final choice will depend on a vote instead of a simple consensus". In WP.it, several forums called *Sondaggio* were set up as an "easy, quick and simple solution for resolving problems" (Langlais 2014:28-30, our translation).

1.1- The main characteristics of WP talk pages

WP talk pages look like traditional online forums and social media, e.g., chats, blogs, and Twitter interactions. They comprise threads about a specific topic composed of user contributions called "posts", which are usually signed with a user alias and a timestamp indicating when the message was posted (e.g., "Hutschi 10:47, 7 Jul 2004 (UTC)" in (1)). When a user wants to discuss a new topic, to ask a question to the community or to make a suggestion for improving the article, he or she has to create a new thread, give it a heading and write a first post. Because there is no publishing process involved, posts may contain nonstandard writing with internet slang and misspellings (e.g. "appoaches" in (1)). Moreover, posts are likely to start with openings such as "hello", end with closures such as "cheers", and contain addressing expressions such as "@Jimbo :" to indicate referencing between these posts, cf. "reply relation" in Lüngen & Herzberg (2019, this volume). Several studies

[4] https://en.wikipedia.org/wiki/User:Jimbo_Wales/Statement_of_principles

have shown that out-of-vocabulary words[5] are fairly rare in WP, with only sparse bits of @addresses, emojis and very rare social media keywords (e.g., U for “you”, “lol”) or internet slang words (cf. Elia 2009, Walton 2009, Myers 2010, Ho-Dac & Laippala 2017).

However, WP talk pages differ greatly from regular social media pages. Three main reasons may be enumerated: First, in WP talk pages, the locutors are all engaged in a common activity, i.e., writing an encyclopedia article. This differs drastically from other social forums where locutors are not coworkers. This common goal and the community feeling should lead to more benevolence between Wikipedians than in common social media, with few threads evolving toward waves of hatred and violence. Nevertheless, a harassment survey published in 2015 by the Wikimedia Foundation indicated that the portion of respondents who had experienced online harassment on WP was approximately the same as on the internet in general (38% on WP compared to 40% in general, according to the Pew Research Center).

Second, WP provides multiple ways for users to interact: discussing on the talk page, commenting on article edits, and writing on their user talk pages, i.e., a talk page associated with their profile page and participating in the various general talk pages, i.e., discussion forums opened in each WP (e.g., the “Teahouse”[6] in WP.en, the “Bistrot”[7] in WP.fr or the “Fragen_von_Neulingen”[8] in WP.de).

Third, talk pages are not composed via dedicated software but use exactly the same process and wikiCode as the articles, where each thread corresponds to one section within which the posts are written as the default text. As a consequence, users are able to edit old posts written by themselves or other users, i.e., to delete or insert content inside, to modify thread heading, etc.

These last two aspects are reasons why talk pages differ from regular online forums. They also have consequences that pose obstacles for building a corpus consisting of whole and coherent discussion threads. Indeed, it may sometimes be challenging to reconstruct a whole discussion thread (cf. Poudat et al. 2017). For example, when a discussion starts on a talk page and ends on a user page or when a suggestion posted on a talk page seems to end abruptly when it is actually an edit in the article that closes the discussion. Another serious type of issue arises when a user edits the wikicode of a thread. A user could at any time insert a new post before an older post and, as a consequence, disrupt the chronological order of the thread. A user could also modify a previous post without changing the timestamp, which makes the timeline somehow artificial. A solution to chronologically reorder posts is to mine the revision history of the talk page either manually (Poudat et al., 2017) or automatically (Hua et al. 2018).

In summary, WP talk pages provide data available under Creative Commons that are fairly well-formed and associated with rich metadata including topics, writer profiles (for those who

[5] OOV words are usually used in NLP in order to characterize noisy data, i.e., data in which there is a large part of words that are not included in standard lexical resources used in NLP (cf. Baldwin et al. 2013).
[6] https://en.wikipedia.org/wiki/Wikipedia:Teahouse
[7] https://fr.wikipedia.org/wiki/Wikipedia:Bistrot
[8] https://de.wikipedia.org/wiki/Wikipedia:Fragen_von_Neulingen

are registered) and shared knowledge (given through the content of the collaborative written article) and that elucidate the collaborative writing process.

1.2 The basic structure of a WP talk page (*tp*)

Basically, a WP talk page (henceforth *tp*) consists of three parts: (1) a header containing information and banners with a variety of disclaimers about the *tp* and the article, (2) the table of contents listing all the threads in the *tp*[9], and (3) the discussion threads and posts. In WP.en and WP.fr, a fourth part occurs at the bottom of the page, listing the topic categories assigned to the *tp*. In WP.de, only articles are associated with topic categories.

Figure 2 provides an extract of the beginning of the *tp* about global warming in WP.en in 2019[10]. After the heading “Talk:Global warming”, two banners inform about the *tp*: the first notices the agreement of the page with a “consistent citation style”, and the second recalls that a “[talk] page is not a forum”. The last two banners provide information about the article’s relevance for dedicated WP projects and/or as educational material.

The right block gives access to the archived pages of the *tp* if an archiving process is performed. Figure 2 shows that the "Talk:Global warming" has 80 archived pages, i.e., pages containing threads that were automatically archived when the current *tp* became too long. As written below the archives list, the WP.en automatically archives “threads older than 30 days”. The archiving process differs depending on the language and may be manual, as in WP.fr. In addition to archived pages, other kinds of *tps* may be manually created and associated with the current *tp*, e.g., "todo list" pages, "npov" pages (i.e., *tp*s dedicated to debate about "**n**eutral **p**oint **o**f **v**iew", one of the basic rules for avoiding bias in articles).

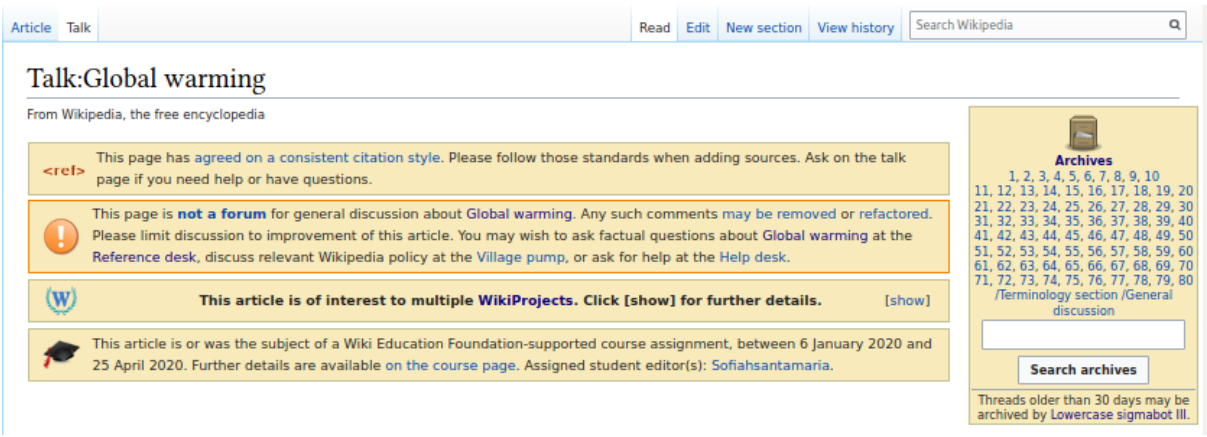


Figure 2: Top of the WP.en page “Talk:Global warming”

The header structure varies considerably according to the language and the *tp*. There are very few banners in WP.de in contrast with WP.en and WP.fr. WP.en and WP.fr

[9] In 2022, the table of content has been moved from the main part of the webpage (between the header and the threads) to the menu (on left for on the computer version).
[10] In 2020, the "Global Warming" and "Climate Change" articles were merged under the single article "Climate Change".

systematically indicate the article quality level in the header, as illustrated in Figure 3, which gives the top of the WP.fr page “Discussion: Réchauffement climatique”.

Figure 3: Top of the WP.fr page “Discussion: Réchauffement climatique”

This figure also signals, before the banner, links to other *tp* that are a French cultural exception corresponding to “parallel” *tps* related to the same article and focusing on dedicated problems such as neutrality problems, article quality level discussions, and to-do lists.

Aside from the banners, a table of content lists all the threads composing the current page. A thread corresponds to a named section created by the user who has posted the first message of the thread. Figure X shows a thread in the WP.en "Talk:Global warming" entitled “*Dispute about what to do about global warming and who should do it - should it be included in the "Public opinion and disputes" section?*”. This thread consists of 6 posts that may be distinguished with the help of indent levels and signatures.

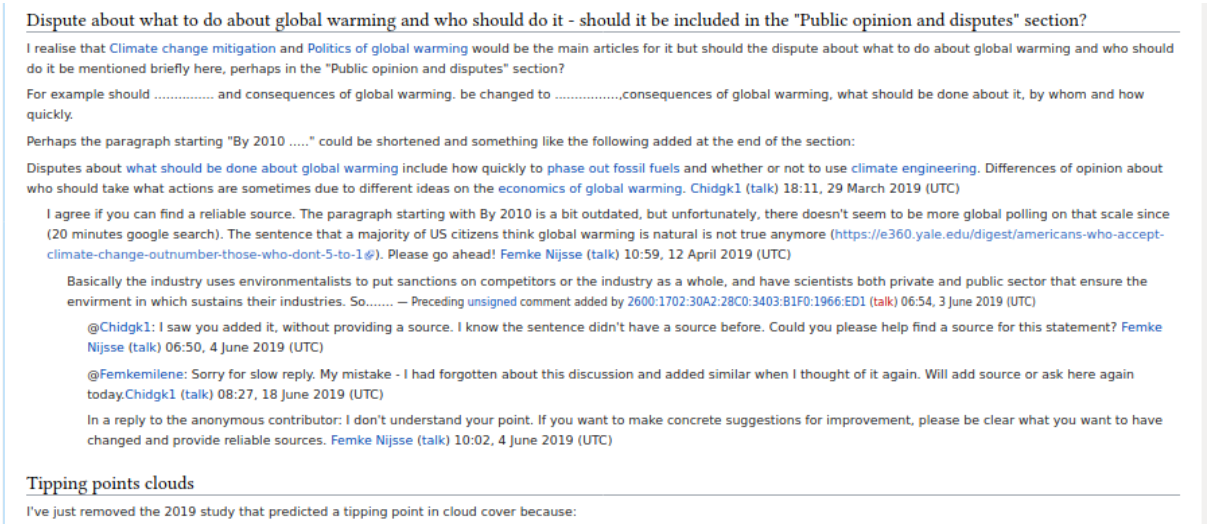


Figure 4: One thread in the WP.en page “Talk:Global warming”

Indent levels normally signal reply relations (cf. Herzberg & Lüngen in this volume). The contributor can also address his message explicitly to a registered user by using @ as seen, for example, in the two penultimate posts in Figure 4 transcribed in (2):

(2)

> *@Chidgk1: I saw you added it, without providing a source. I know the sentence didn't have a source before. Could you please help find a source for this statement? Femke Nijsse (talk) 06:50, 4 June 2019 (UTC)*
> *@Femkemilene: Sorry for slow reply. My mistake - I had forgotten about this discussion and added similar when I thought of it again. Will add source or ask here again today.Chidgk1 (talk) 08:27, 18 June 2019 (UTC)*

In this example, the contributors signaled their reply to the previous message by using a reply-to template without incrementing the indent level[11]. Another clue ensuring that these are two different posts and not one single post is the final signature. A signature is inserted by the user at the end of the post, and its content is automatically generated (cf. https://en.wikipedia.org/wiki/Wikipedia:Signatures). When a user forgets to insert it, it may be automatically generated as for the 3rd post in Figure 4. Signatures include the timestamp and the user identity, i.e., the Wikipedian's registered name or the user IP in case of an unregistered user.

The thread shown in Figure 4 is a good example of an asynchronous, postedited timeline with varying time intervals between rounds and potential inconsistencies. Example (3) provides an abbreviated version of this thread with only the users' identities, timestamps and the beginning of each message.

(3)

- **Chidgk1**: *I realise that …* (18:11, 29 March 2019)
- **Femke Nijsse**: *I agree if…* (12 days later -- 10:59, 12 April 2019)
- **Anonymous contributor**: *Basically …* (almost 2 months later -- 06:54, 3 June 2019)
- **Femke Nijsse**: *@Chidgk1: I saw you added…* (the day after -- 06:50, 4 June 2019)
- **Chidgk1**: *@Femkemilene: Sorry for slow reply…* (2 weeks later -- 08:27, 18 June 2019)
- **Femke Nijsse**: *In a reply to the anonymous contributor: I don't understand…* (2 weeks before, when addressing to Chidgk1, -- 10:02, 4 June 2019)

Example (3) provides a good example of the temporal inconsistency of the thread given in Figure 4. When looking at the timestamps in the post, we can understand that the last message was posted the 4th of June "in reply to [an] anonymous contributor" that wrote "Basically…" the day before (the 3rd of June). However, Chidkg1 edited the talk pages on the 18th of June by inserting a message ("Sorry for the slow reply…") between the two messages posted the 4th of June by Femke Nijsse. As a consequence, the interaction through the last three rounds sounds very difficult to reconstruct a posteriori (cf. Poudat et al. 2017).

Despite this temporal inconsistency, the thread in Figure 4 looks like a regular conversation with someone launching a discussion and one or several other users reacting to it. Nevertheless, this is not the usual case on WP. A large number of threads consist of only one post or several posts from one single user, as in the minimalist WP.de page "Diskussion:Tennenbach", which contains only two words without any thread heading or signature (Figure 5). Tanguy, Poudat and Ho-Dac, in this volume, provide a quantitative

[11] WP proposes an enormous amount of templates for helping users write articles and talk pages. WP templates are described below. The *reply-to template* was introduced in the WP.en article https://en.wikipedia.org/wiki/Template:Reply_to.

overview of the interaction pattern distribution in the EFG WikiCorpus, with approximately half the threads composed solely of a single post.

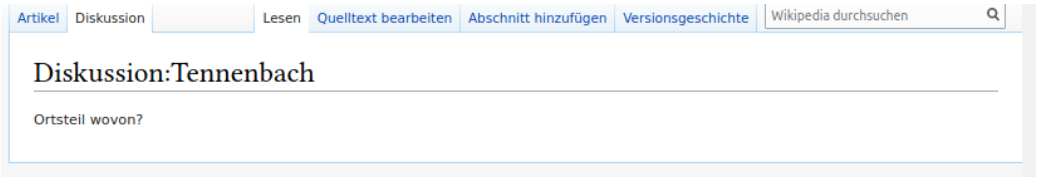


Figure 5: The entire WP.de page “Diskussion:Tennenbach”

Section 4 gives a quantitative overview of the talk pages in the EFG WikiCorpus.

1.3 Talk page encoding: The TEI CMC-core schema

The CMC-core (Beißwenger and Lüngen 2020) proposes a schema for encoding CMCs (Computer Mediated Communications) according to the Text Encoding Initiative (TEI) guidelines. The TEI is a community-defined encoding model for texts in the humanities, including speech and language corpora. It is widely used and accepted in the humanities and hence can be viewed as a de facto standard in this community. The CMC-core schema aims at integrating heterogeneous CMC sources under a common model and, as a consequence, facilitating open-corpora sharing with the research community (Beißwenger et al. 2017). It was involved in the various already defined TEI elements, such as those for encoding participants and turns speaking (as in Performance Texts, TEI-P5 Chapter 7) or for encoding signatures (as in Manuscript Description, TEI-P5 Chapter 10).
The brand-new textual element proposed in the CMC core is the <post> element for delimiting user turn taking and recording an anonymous @who and a @when-iso location[12]. A new class attribute @creation was also proposed for distinguishing posts written by humans or by machines (e.g., bots). The other elements and attributes used for encoding CMCs features were already defined in TEI-P5 guidelines, e.g., <teiHeader> vs. <text> for recording metadata vs. textual content, for delimiting sections, i.e., threads, <head> for thread headings, <signed> for signatures including the timestamp (<date>) and eventually the user <name>, and <ref> for hyperlinks.

As an example, the thread opening post in (1) will be TEI CMC-core encoded as in (4):

(4)

```
<div type="thread" xml:id="i.66242_5">
  <head>Chomsky and Corpus Linguistics</head>
  <post indentLevel="0" mode="written" when-iso="2004-07-07T10:47+00" who="WU00046413" xml:id="i.66242_5_1">
    <p> The article states: "The approach runs counter to Noam Chomsky's view that real language is riddled with performance-related errors, thus requiring careful analysis of small speech samples obtained in a highly controlled laboratory setting. "</p>
    <p> When did Chomsky say this and where? Do the two appoaches contradict each other or do they complete each other?</p>
    <p>--<signed type="signed"> <ref target="https://en.wikipedia.org/wiki/User:Hutschi"><name
```

[12] Time location must be normalized according to ISO 8601 based on the Coordinated Universal Time (UTC). The list of TimeZone abbreviations and UTC offset correspondences may be found here : https://www.timeanddate.com/time/zones/.

```
full="yes">Hutschi</name></ref><date>10:47, 7 Jul 2004 (UTC)</date></signed></p>
    </post>
[...]
</div>
```

2- Building the EFG WikiCorpus

Because WP use approximately the same technology and policies in every language, common tools could be developed for all language versions. For example, the Contropedia platform gives access to the controversies that concern WP users by analyzing the article's revision history (text spans frequently edited and reverted) and talk pages (Laniado 2011, Borra et al. 2015 and Laniado et al. in this volume). The Convokit toolkit (Chang et al. 2020) provides natural language processing tools for analyzing the discussions occurring in the web, including WP talk pages. Such tools are very useful for the field of politeness strategies, conflict analysis and conversational failure (cf. Zhang et al. 2018 and Chang & Danescu-Niculescu-Mizil, 2019). Recently, Hua et al. (2018) proposed a pipeline for "reconstructing a complete and structured history of the conversational process in Wikipedia talk pages". This pipeline provides the "WikiConv" multilingual corpus (English, Chinese, German, Greek and Russian) composed of talk pages and information about the authors' editing actions. The EFG WikiCorpus we present in this chapter is, as the WikiConv corpus, a comparable corpus of discussions in English, French and German encoded according to the TEI CMC-core schema.

Two methods may be used for building corpora from Wikipedia: starting from the Wikipedia website, crawling all pages and parsing the html code for extracting relevant content (e.g., Zesch 2008, Hua et al. 2018, Mitrevski 2020, Kraif in this volume); or starting from the Wikipedia archives, downloading an official dump and parsing the wikicode and retrieving metadata from the SQL (e.g., Lainado 2011, Laniado et al. in this volume, Margaretha & Lüngen 2014, Ho-Dac et al. 2017, Linguatools 2018).

While the first technique seems to be easier for extracting the current content of articles, the second technique seems more adapted if we want to benefit from all available metadata (e.g., thematic portals and categories, article quality), interlingual information and templates that may encode relevant information about user intention (e.g., reply to and addressing templates). This is why we choose the second technique for building the EFG WikiCorpus.

WP archives for each language are regularly provided as database dumps on the web by the Wikimedia foundation. These dumps contain all WP pages, including articles, talk pages, user pages, and user talk pages, in two versions: the current version and the version with all revisions that have been made on the pages since the creation of the page. The dumps also contain databases providing, for example, multilingual links between articles, categories and keywords. Hence, the WP archives provide all the data needed for building comparable corpora of both articles and discussions.

3.1- Searching for relevant content in the WP archives and the wikicode

The Wikimedia Foundation regularly publishes official dumps that provide “a complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML. A number of raw database tables in SQL are also available. These snapshots are usually provided twice a month" (cf. https://dumps.wikimedia.org/).

The main file used for building linguistics corpora from the Wikipedia archives is the XML file gathering together all the current pages (without edit history) of a WP version and entitled *xx*wiki-*yyyymmdd*-pages-meta-current.xml.bz2, where xx indicates the language (e.g., dewiki, enwiki, frwiki) and the yyyymmdd the backup date. The database of this volume is based on the English, French and German WP versions dated August 1st 2019:

- dewiki-20190801-pages-meta-current.xml,
- enwiki-20190801-pages-meta-current.xml,
- frwiki-20190801-pages-meta-current.xml.

In addition, SQL files providing Wiki interlanguage link records are necessary for aligning articles, portals and category metadata in the three languages: *xx*wiki-*yyyymmdd*-langlinks.sql.

The *xx*wiki-*yyyymmdd*-pages-meta-current.xml files provide an XML format structured where each current page is embedded in an XML element `<page>`[13] containing a header encoded in an XML format and the body of the page encoded in the wikicode format.

The wikicode format is the markup language (also known as wiki markup or wikitext) used by the MediaWiki software. It consists of a light markup to format page layout such as links, headings, lists, boldface and italics. This light markup is strongly enlarged by a wide open-ended list of templates. WP templates are used for embedding a wide variety of recurring content in a page, such as special characters (e.g., the WP template {{lambda}} will be “transcluded”, i.e., transformed into the character lambda symbol “λ” on the html webpage and `{{s-|XV}}` into “XV^{e} siècle”), footnotes, information banners, warnings in the header, etc. WP templates may also point to small scripts used for computing information such as time location[14]. Apart from these templates that are mainly used in articles, some dedicated templates occur in talk pages for inserting emojis and, more specifically, for indicating interactions such as a reply to, e.g., `{{reply to|Chidgk1}}` will be “transcluded” into “@*Chidgk1*” on the html webpage with a direct link to the user WP page, a vote, an unsigned post, the fact that a job has been done, etc.

Figure 6 illustrates the beginning and the end of the `<page>` relative to the article “Global Warming” (on left) and the talk page “Talk:Global Warming” (on right).

[13] All extracts from the dump are written with a monospaced font.

[14] For example, the template `{{CENTURY|YYYY}}` returns the calendar century number for the Georgian year YYYY.

```
<page>
  <title>Global warming</title>
  <ns>0</ns>
  <id>5042951</id>
  <revision>
    <id>908741870</id>
    <parentid>908736853</parentid>
    <timestamp>
      2019-07-31T17:44:40Z
    </timestamp>
    <contributor>
    <username>AnomieBOT</username>
    <id>7611264</id>
  </contributor>
[...]
  <text xml:space="preserve">
{{...}}
{{Use      British      English      Oxford
spelling|date=June 2019}}
{{...}}

'''Global warming''' is the current long-
term rise [...]
[...]

{{Global warming|state=expanded}}
{{Human impact on the environment}}
{{...}}
[[Category:Global warming| ]]
[[Category:Climate change]]
[[...]]
   </text>
 </revision>
</page>
```

```
<page>
  <title>Talk:Global warming</title>
  <ns>1</ns>
  <id>454409</id>
  <revision>
    <id>908794776</id>
    <parentid>908794735</parentid>
    <timestamp>
      2019-08-01T01:12:26Z
    </timestamp>
    <contributor>
      <username>NewsAndEventsGuy</username>
      <id>14536509</id>
    </contributor>
[...]
    <text xml:space="preserve">
{{Skip to talk}}
{{Talk header|noarchives=yes}}
{{Vital
article|level=3|topic=Science|class=FA}}
{{...}}
{{Not a forum}}
{{WikiProject
Arctic|class=FA|importance=high}}
{{WP1.0|v0.5=pass|class=FA|category=Geograp
hy|coresup=yes|VA=yes|WPCD=yes}}
}}
{{...}}

== Dispute about what to do [...]==
I realise that [...]
   </text>
 </revision>
</page>
```

Figure 6: WP.en dump extract

Metadata contained in the header are not displayed on the html page and provide information about the page status: its title (`<title>`), the namespace it belongs to (`<ns>`), its id in the database (`<id>5042951</id>` for the article and `<id>454409</id>` for the talk page in Table 1), the date of this current version, i.e., the last revision date (`<timestamp>`) and the last contributor, if the page has been redirected, translated, etc.

3.2. Extracting talk pages and TEI encoding metadata

Building a multilingual corpus of talk pages must be a three-way process: parsing the langlinks.sql database to obtain multilingual links, parsing the article to obtain thematic metadata, and parsing the talk page to obtain the content of the discussion and additional metadata.

The langlinks.sql database contains triplets indicating for all pages the title of its counterpart in all available other languages. For example, the WP.en langlinks.sql file contains the two triplets (`5042951,'fr','Réchauffement climatique'`) and (`5042951,'de','Globale Erwärmung'`). These triplets indicate that the WP.en page with the id `5042951` (the article

“Global Warming”) corresponds to a page entitled "Réchauffement climatique" in French (`'fr'`) and "Globale Erwärmung" in German (`'de'`). We parsed the langlinks.sql files of WP.en, WP.fr and WP.de and extracted all EFG page links.

Next, we used namespaces to distinguish the different types of pages recorded in the dump. Namespace 0 corresponds to article pages (`<ns>0</ns>`), 1 to talk pages (<ns>1</ns>), 2 to user pages, 14 to Wikipedia Category pages, etc. The namespace is also indicated in the title pages, except for articles: `<title>Talk:Global Warming</title>`, `<title>User:Jimbo_Wales</title>`, `<title>Category:Global Warming</title>`, etc.

In article pages (`<ns>0</ns>`), the `<text>` element usually starts with templates `{{...}}` informing about the article lifecycle, e.g., if the page has been redirected, if the article is mostly a translation or a copy, and if, as in Figure 6, the article has been spell-checked (`{{Use British English Oxford spelling|date=June 2019}}`). Relevant keywords `{{...}}`, thematic portal `[[Portal:XXX]]` and category `[[Category:XXX]]` memberships are listed at the end of the `<text>` element.

We used the portals, categories and keywords for structuring the EFG WikiCorpus according to topics. These thematic gatherings are organized hierarchically with at the first level approximately 10 main Portals (see Table 1 below) divided into sub Portals and Categories. For example, the WP.en article "Corpus Linguistics" is tagged as being part of the main Portal "Society and social sciences", the sub Portal "Linguistics" and the categories "Corpus linguistics", "Applied linguistics", "Discourse analysis", "Linguistic history", and "Linguistic research". Keywords are independent of this hierarchy.

Because each WP has its own nomenclature, many pages, whether about a topic, a portal or a category, have no official counterpart in the other languages, i.e., no links to WP.en, WP.fr and/or WP.de. Table 1 lists the main portal names available in the EFG WikiCorpus and proposes a potential alignment based on our interpretation of the topics the portals seem to cover.

13 WP.en main Portals	11 WP.fr main Portals	8 WP.de main Portals
Geography and places	Géographie	Geographie
History and events	Histoire	Geschichte
Religion and belief systems	Religion	Religion
Technology and applied sciences Mathematics and logic Nature and sciences Philosophy and thinking	Technologie Sciences	Technik Wissenschaft
Society and social sciences	Société Politique	Gesellschaft

Culture and the arts	Art Sport	Kunst und Kultur Sport
Health and fitness	Médecine	
Human activities	Loisir	
People and self		
Reference		

Table 1: EFG WikiCorpus thematic portals[15]

Multilingual links between topics may also be established between the categories by using the triplets related to the id of the category page in the langlinks.sql file. In the EFG WikiCorpus, only the links toward the WP.en category pages have been encoded (see below).

After the thematic information was extracted from the article page, the talk page could be parsed. As in article pages, the `<text>` element in talk pages starts with templates informing about the talk page status and about the article status. This information is displayed as banners in the header of the html version of the talk page. In the WP.en and WP.fr talk pages, these templates always indicate the article quality and its significance according to specific thematic portals and categories. In Figure 6, the template `{{Vital article|level=3|topic=Science|class=FA}}` indicates that the "Global warming" article is vital according to the "science" portal and of FA quality (i.e., Featured Article). The next template `{{WikiProject Arctic|class=FA|importance=high}}` indicates that the article is highly significant for the WikiProject "Arctic". Other templates may be used for displaying warning messages and reminders such as `{{Not a forum}},` which displays a banner reminding users that Wikipedia is not a forum, or `{{calm}},` which prevents contributors to potential conflicts in the discussion.

In the target representation of the EFG corpus, all extracted metadata and multilingual links are encoded in the <teiHeader>, i.e., the related article page and its ID, interlingual links, thematic portals and categories, and keywords. When available, the counterparts in the other languages are indicated in the <relatedItem> and <classCode> TEI elements. For portals and categories, we decided to link WP.fr and WP.de portals and categories to the English-aligned when counterparts are recorded in the langlinks.sql database. Example (5) gives an extract of the resulting <teiHeader> of the WP.fr talk page about Global Warming. The <relatedItem> elements encode the links to the article and its id in the local language (@n="25425" and @targetLang="fr") and to its counterparts in the two other languages. The <classCode> elements encode the article quality (here "B") and the link to the Portals and

[15] https://en.wikipedia.org/wiki/Portal:Contents/Portals, https://fr.wikipedia.org/wiki/Portail:Accueil, https://de.wikipedia.org/wiki/Portal:Wikipedia_nach_Themen

Categories in the local language and in English. Only main Portals are displayed in (5), and all URLs are replaced by the placeholder term “URL” for readability.

(5)

```
<teiHeader>
<fileDesc>
  <titleStmt>
    <itle>Discussion:Réchauffement climatique</title>
  </titleStmt>
  …
```

```
  <sourceDesc>
    …
    <biblStruct>
      …
```

```
      <relatedItem type="langLink">
        <ref target="URL" targetLang="de">Globale Erwärmung</ref>
      </relatedItem>
```

```
      <relatedItem type="langLink">
        <ref target="URL" targetLang="en">Global warming</ref>
      </relatedItem>
      <relatedItem type="articleLink">
        <ref n="25425" target="URL" targetLang="fr">Réchauffement climatique</ref>
      </relatedItem>
    </biblStruct>
  </sourceDesc>
</fileDesc>
<profileDesc>
```

```
      …
```

```
  <textClass>
    <classCode scheme="https://fr.wikipedia.org/wiki/Projet:Évaluation">
      <ref target="URL">B</ref>
    </classCode>
    <classCode scheme="https://fr.wikipedia.org/wiki/Portail:Accueil">
      <ref target="URL" targetLang="fr">Portail:Politique</ref>
      <ref target="URL" targetLang="fr">Portail:Sciences</ref>
      <ref target="URL" targetLang="fr">Portail:Géographie</ref>
      <ref target="URL" targetLang="fr">Portail:Environnement</ref>
      <ref target="URL" targetLang="fr">Portail:Écologie</ref>
```

```
      <ref target="URL" targetLang="fr">Portail:Énergie</ref>
```

```
      <ref target="URL" targetLang="fr">Portail:Météorologie</ref>
      …
    </classCode>
    <classCode scheme="https://fr.wikipedia.org/wiki/Catégorie:Accueil">
      <ref target="URL" targetLang="fr">Catégorie:Thermodynamique atmosphérique</ref>
      <ref target="URL" targetLang="fr">Catégorie:Changement climatique</ref>
      <ref target="URL" targetLang="fr">Catégorie:Effet de serre</ref>
      <ref target="URL" targetLang="fr">Catégorie:Catastrophe environnementale</ref>
    </classCode>
    <keywords>
      <term>Monde polaire</term>
      <term>Énergie</term>
      …
    </keywords>
    <classCode scheme="https://en.wikipedia.org/wiki/Category:Contents">
      <ref target="URL" targetLang="en">Category:Atmospheric thermodynamics</ref>
      <ref target="URL" targetLang="en">Category:Climate change</ref>
      <ref target="URL" targetLang="en">Category:Global warming</ref>
      <ref target="URL" targetLang="en">Category:Environmental disasters</ref>
    </classCode>
    <classCode scheme="https://en.wikipedia.org/wiki/Wikipedia:Contents/Portals">
```

```
      <ref target="URL" targetLang="en">Portal:Politics</ref>
      <ref target="URL" targetLang="en">Portal:Science</ref>
      <ref target="URL" targetLang="en">Portal:Geography</ref>
      <ref target="URL" targetLang="en">Portal:Environment</ref>
      <ref target="URL" targetLang="en">Portal:Ecology</ref>
      <ref target="URL" targetLang="en">Portal:Energy</ref>
    </classCode>
  </textClass>
</profileDesc>
</teiHeader>
```

3.3- Parsing the wikicode and TEI CMC-core encoding

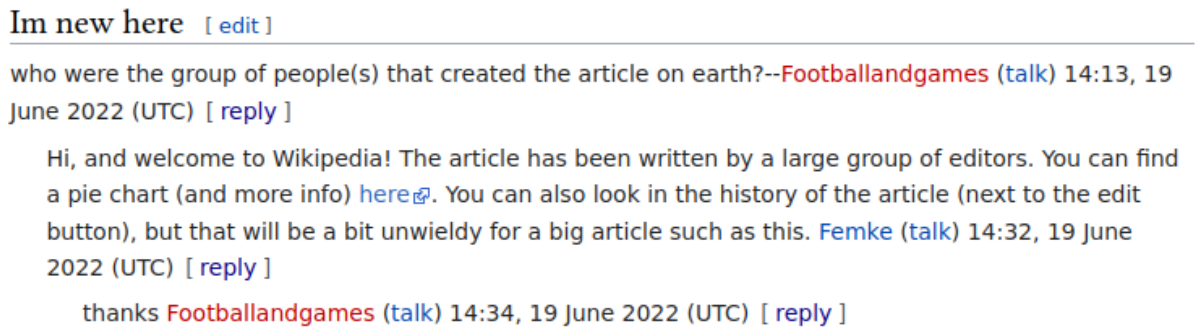
As stated previously, the <text> element encapsulates the page content encoded in the wikicode syntax. The two next sections list the global structure of the textual content of a talk page through its wikicode and the resulting TEI encoding.

3.2.1- The global content structure of a talk page

The structure of a talk page is fairly simple. Each thread corresponds to a section that starts with a heading. As a result, starting a new thread consists of writing a subsection heading by using the wikicode syntax, i.e., *== Subsection Heading ==*[16] followed by one or more paragraphs constituting the first post of the thread. Once the message is complete, the user must sign his post by using a template signature (see next section).

Most of the time, when a user posts a new message in an existing thread, he or she simply has to insert a new post with an indent level (**:** in wikicode). As a result, a thread composed of 3 posts in which A asks for something, B answers and A expresses thanks is displayed on the talk page as in (6), which is extracted from the Earth WP.en talk page (the associated wikicode is given below).

(6)




```
== Im new here ==

who were the group of people(s) that created the article on earth?--
[[User:Footballandgames|Footballandgames]] ([[User talk:Footballandgames|talk]]) 14:13, 19
June 2022 (UTC)
```

[16] The section heading *== Section Heading ==* is dedicated to the page title and should not be used in the body of a page.

```
: Hi, and welcome to Wikipedia! The article has been written by a large group of editors. You can find a pie chart (and more info) [https://xtools.wmflabs.org/articleinfo/en.wikipedia.org/Earth here]. You can also look in the history of the article (next to the edit button), but that will be a bit unwieldy for a big article such as this. [[User:Femkemilene|Femke]] ([[User talk:Femkemilene|talk]]) 14:32, 19 June 2022 (UTC)

::thanks [[User:Footballandgames|Footballandgames]] ([[User talk:Footballandgames|talk]]) 14:34, 19 June 2022 (UTC)
```

Apart from these fundamental layout features, other layout features, such as lists, italics, bold, and tables, may be used. Nevertheless, these other layout features occur more frequently in article pages than in talk pages.

(6) also shows templates, i.e., text spans delimited in the wikicode with double brackets ([[ ]]). Templates are Wikicode tools designed to facilitate writing by replacing frequent predefined strings (e.g., date, emoticon, user signature) by simple character combinations. For example, typing 4 tildes (~~~~) will generate a signature composed of the user identity (its pseudonym if registered or its IP if not) and the timestamp. The result is, for example, "*Femke (talk) 14:32, 19 June 2022 (UTC)*" at the end of the second post in (6), corresponding to "`[[User:Femkemilene|Femke]] ([[User talk:Femkemilene|talk]]) 14:32, 19 June 2022 (UTC)`" in the Wikicode.

3.2.2- Structuring and encoding the threads into posts

Structuring the EFG WikiCorpus into posts has been processed automatically with a rules-based method using two main complementary rules: if a signature ends a paragraph and/or if a change of indent level is indicated with one or more colons (:) at the beginning of a new paragraph.

The user signature is the most relevant and reliable cue for detecting post endings. Unfortunately, many variations have been observed in the signatures (even if signatures are usually automatically generated by the template), and a signature is not always the end of the current post considering that users may insert a new post in a previous post or erase the signature. Example (7) illustrates a thread[17] beginning with an unsigned post that could only be delimited from the next signed post because of a change of indent level (`::`). The opening of the second post clearly demonstrates that unsigned posts compromise post identification and, as a consequence, discussion fluency.

(7)

```
== First sorry, then a suggestion ==

Hello.

Sorry if I caused unwanted trouble with the "Creationism2" template. My purpose was to remove the box from the Flat Earth article -- not to delete the template itself. Sorry if the latter happened.
```

17 https://en.wikipedia.org/wiki/Talk:Modern_flat_Earth_beliefs/Archive_1#First_sorry,_then_a_suggestion

```
Now, because it seems I don't have the computer skills myself, I strongly suggest someone to remove the box if I didn't succeed in the proper way.

The reason for this is quite clear. [...] Although some models combine better with the indirect evidence, no-one has to play fool and think contrary of what is seen today.

::The above was unsigned, maybe it is very old. I have to say that although it is questionable whether a neutral reliable source can be found that would say all creationism is as bad as the flat earth, you would probably find many that compared the two and may even find some that equates the more extreme forms to flat Earthers. [...] I do like that the flat Earthers have the world as a round surface; that helps with some of them. [[User:Rifter0x0000|Rifter0x0000]] ([[User talk:Rifter0x0000|talk]]) 11:01, 21 July 2010 (UTC)
```

Three kinds of signatures are distinguished in the EFG WikiCorpus and encoded using the @type attribute of the <signed> TEI element: @type="signed" indicates that the post was explicitly signed by a registered user using a user signature template (e.g., ~~~~); @type="unsigned" indicates that the post was marked by either a registered or unregistered user using the Unsigned or Help template; @type="user_contribution" indicates that the corresponding posting was marked using a `[[Special:Contributions/IP]]` link (e.g., by an unregistered user). The user name indicated in the signature was encoded with the <ref> (linked to the user WP page) and <name> TEI elements and was also given in an anonymized version in the @who attribute. The timestamp is encoded in a <date> element, and the ISO 8601 time location is indicated in the @when-iso attribute in the <post> element[18]. Colon(s) indicating the indent level have been removed, and the indent level is explicitly given in the @indentLevel attribute.

Example (8) shows the TEI CMC-core encoding of (7). As indicated in the @xml:id attribute, this thread is the 55th of the talk page identified as 19583719 in the WP dump. The first post indentLevel is "0" in contrast to the second post indentLevel, which is "2". In this thread, there is no post at the first indentLevel because the user Rifter0x0000 inserted two colons instead of one at the beginning of his reply. Because the first post was unsigned, the value of the @who attribute is the one dedicated to anonymous users, i.e., "`WU00000000`". The @who value of the second post was the one automatically associated with the user called Rifter0x0000, i.e., `WU00017020`. Because the timestamp was automatically inserted when the user signs his post, the first post has no @when_iso attribute as opposed to the second post.

(8)

```
<div type="thread" xml:id="i.19583719_55">
   <head>First sorry, then a suggestion</head>
   <post indentLevel="0" mode="written" who="WU00000000" xml:id="i.19583719_55_1">
      <p> Hello.</p>
     <p> Sorry if I caused unwanted trouble with the "Creationism2" template. My purpose was to remove the box from the Flat Earth article -- not to delete the template itself. Sorry if the latter happened.</p>
     <p> Now, because it seems I don't have the computer skills myself, I strongly suggest someone to remove the box if I didn't succeed in the proper way.</p>
```

[18] ISO 8601 time locations are based on the Coordinated Universal Time (UTC) e.g. the main part of German and French posts are time stamped acc. to the Central European Time (CET) or Central European Summer Time (CEST) that corresponds to UTC+01 or UTC+02. The list of TimeZone abbreviations and UTC offset can be found at https://www.timeanddate.com/time/zones/.

```
     <p> The reason for this is quite clear. [...] Although some models combine better with
the indirect evidence, no-one has to play fool and think contrary of what is seen
today.</p>
   </post>

 <post indentLevel="2" mode="written" when-iso="2010-07-21T11:01+00" who="WU00017020"
xml:id="i.19583719_55_2">
     <p>The above was unsigned, maybe it is very old.  I have to say that although it is
questionable whether a neutral reliable source can be found that would say all creationism
is as bad as the flat earth, you would probably find many that compared the two and may
even find some that equates the more extreme forms to flat Earthers. [...]  I do like that
the flat Earthers have the world as a round surface; that helps with some of them.  <signed
type="signed"><ref           target="https://en.wikipedia.org/wiki/User:Rifter0x0000"><name
full="yes">Rifter0x0000</name></ref><date>11:01, 21 July 2010 (UTC)</date></signed></p>
   </post>
</div>
```

A prototype of this post encoding method has been evaluated in a previous study (cf. Ho-Dac and Laippala 2017). Most of the observed errors have been corrected even if some tricky wikicode segments still remain difficult to parse correctly, as in (9)[19], where @when_iso is outside of possible time stamps but has been manually written by the user.

(9)

```
Storylines cleanup

All the subheadings under the "Storylines" section need serious cleanup; not every
subheading is a storyline in its own right and should be merged together. The earlier
storylineás are pretty well-kept and appropriate, but towards the end the subheadings
increase unnecessarily breaking up entire storylines into subheadings of only a few
episodes (or less). •97198 talk 17:28, 33 June 2007 (UTC)
```

Another type of segmentation error may be caused by the fact that everything in the Wikipedia world can be postedited. A post could be inserted before a post published previously[20]. Anytime, an author could edit an existing post by inserting new content into it. Such postedition will cause an error in the segmentation, i.e., the two parts of the post preceding and following the insertion will be considered as two posts[21]. As a consequence, it is sometimes impossible to properly understand the content structure without investigating the revision history (cf. Hua et al. 2018).

3.2.3- Templates and special features

Aside from content structure encoding, other features have been encoded in the EFG WikiCorpus. These special features are mainly wikicoded as templates. The most frequent and relevant templates used in talk pages are those that are also used for writing article

[19] extracted from https://en.wikipedia.org/wiki/Talk:Josh_Ashworth

[20] An example of such insertion could be read in the page "Talk:Harvey_Weinstein/Archive_1" where a post was inserted on the 13th of October before a post published on the 12th of October: https://en.wikipedia.org/wiki/Talk:Harvey_Weinstein/Archive_1#Weinstein's_Wife_(Georgiana_Chapman)_Announced_her_Separation_from_Him

[21] An instance of this can be found in the page "Talk:Friedensreich_Hundertwasser" where a bot, called InternetArchiveBot, inserted on 22 January new content into a post it wrote in October 2017, causing an error: https://en.wikipedia.org/wiki/Talk:Friedensreich_Hundertwasser#External_links_modified

pages, e.g., hyperlinks, notes and quotations. However, some templates are dedicated to talk pages and interacting, such as those for explicitly addressing a message to a registered user (e.g., reply to, greetings, mention of a registered user), those for inserting emojis, and those for special actions such as indicating a conflict, censor a text span, vote (pro vs. con), and check to do lists. The textual content of the templates was marked differently depending on the template purpose:

- addressing templates are encoded with the TEI:ref element that marks up the contributor to whom the template is addressed by using @type and @rend attributes for indicating the speech act (e.g., calling, greeting, welcoming) and the visual rendering (e.g., reply to, ping, Hi, bonsoir, Danke);
- emphasis and quote templates are encoded with the TEI:emph and TEI:quote elements, respectively, by using @type and @rend attributes for indicating the emphasis type and the visual rendering;
- emojis templates are encoded with a TEI:desc element embedded in a TEI:figure element.

(10) shows a post extracted from the page "Talk:Anorexia nervosa/Archive 4" containing a call to the user "SandyGeorgia" and the emoji "wink" templates. The encoding in TEI:XML format is given below.

(10)

No problem, SandyGeorgia. As a general rule, I don't go hunting for sources at Google Books. However, I do try to clean up these refs when I see 'em. —Shelley V. Adams ‹blame credit› 00:42, 10 April 2015 (UTC)

```
<post indentLevel="2" mode="written" when-iso="2015-04-10T00:42±00" who="WU00042432" xml:id="i.46179280_14_6">
  <p>
    No problem, <name creation="template" type="user">SandyGeorgia</name>. As a general rule, I don't go hunting for sources at Google Books. However, I do try to clean up these refs when I see 'em. <figure creation="template" rend="emoji" type="emoji"><desc type="template">wink</desc></figure> –› <signed type="unsigned"><date>00:42, 10 April 2015 (UTC)</date></signed>
   </p>
</post>
```

4- The resulting EFG WikiCorpus

The resulting EFG WikiCorpus is composed of more than 3.3 million *tp*s and 2 billion words and almost 10 million threads and 30 million posts[22]. Table 2 provides the number of all article and talk pages in the August 1st 2019 de/en/fr dumps and the proportion of the *tp*s included in the EFG WikiCorpus. The number of talk pages differs from the number of article pages because there are articles for which no conversation has been launched with the consequence that no associated *tp* exists[23] and because there are articles for which several *tp*s have been created. For example, the article about Global Warming has, in August 2019,

[22] The EFG_WikiCorpus is available on Ortolang: https://hdl.handle.net/11403/efg-wikicorpus
[23] See for example: https://en.wikipedia.org/wiki/Jacques_(band)

91 *tp*s: the current *tp*, 78 archived *tp*s[24] (entitled "Talk:Global Warming/Archive XX") and 12 parallel *tp*s (e.g. "Talk:Global warming/List of archives", "Talk:Global warming/to do", etc.).

lang.	#article pages	#*tp*s	#*tp*s in the EFG WikiCorpus	%*tp*s from the dump included in the EFG WikiCorpus
E	14,856,106	7,903,148	2,025,888	26
F	3,729,677	1,852,689	266,699	14
G	3,920,295	769,091	713,485	93
EFG	22,506,078	10,524,928	3,006,072	29

Table 2: Number of article and talk pages (*tp*s) in the August 1st, 2019, de/en/fr dumps and % of the *tp*s included in the EFG WikiCorpus composed of the current pages, archived pages and npov pages containing at least 1 post (or thread) and 2 words.

As shown in Table 2, less than 30% of the *tp*s archived in the three dumps are in the EFG WikiCorpus, with drastic variations from WP.de to WP.en or WP.fr. This proportion is partly because the WP dumps include empty *tp*s such as those that consist only of a redirect to another one (e.g., the page "Global Warming" with a capital “W”). The other reason explaining the weak proportion is that we selected *tp*s that contain at least 1 post and 2 words and that are current pages, archived pages or "npov" pages for being part of the EFG WikiCorpus.

4.1- Quantitative overview of the talk page content

Table 4 provides a quantitative overview of the EFG corpus in terms of thread and post segmentation and encoding. The main part of the EFG WikiCorpus is the current *tp*s, with only 4% of archives or npov *tp*s. Approximately 40% of the EFG WikiCorpus are only composed of one single post (# single post talks). In other words, 60% of the EFG WikiCorpus are real discussions with at least two messages. This large number of single post talks skews the median number of posts per *tp* toward 2 posts, even if outsider talks may be composed of thousands of posts (up to 2,885 for the WP.en page "Talk:Waterboarding/Archive 7"[25]).

The E WikiCorpus subcorpus is approximately ten times larger than the F WikiCorpus, with the G WikiCorpus in between. Whether in terms of the number of *tp*s, threads, posts or words, the E WikiCorpus represents approximately 75% of the EFG WikiCorpus. The talks are shorter in F WikiCorpus with an average of 6 posts per *tp*s against 9 in E WikiCorpus and G WikiCorpus, but posts are longer in F_WikiCorpus with 87 words per post.

	E WikiCorpus	F WikiCorpus	G WikiCorpus	EFG WikiCorpus
Talk pages	2,025,888	266,699	713,485	3,006,072

[24] Archived *tp*s the *tp* https://en.wikipedia.org/wiki/Talk:Global_warming/Archive_index
[25] https://en.wikipedia.org/wiki/Talk:Waterboarding/Archive_7

Archive *tp*s[26] (%)	57,432 (3%)	4,836 (2%)	23,052 (3%)	85,320 (3%)
#threads	6,636,783	608,857	2,121,852	9,367,493
#posts	20,025,945	1,832,416	6,938,168	28,796,529
Posts/tp : max median mean	2,885 2 9.2	1,976 2 6.9	1,230 2 9.1	2,885 2 9
#single post talks (%)	768,985 (38%)	131,658 (49%)	286,163 (40%)	1,186,806 (39%)
#words[27]	1,448,411,901	146,230,896	454,737,723	2,049,380,520
Words/post : mean	78	87	70	77

Table 4: Quantitative overview of the EFG WikiCorpus.

4.2- Metadata overview and multilingual alignments

As explained in Section 3.2, each discussion is associated with rich metadata (authorship, timestamps, article quality, thematic portals, categories and keywords) and interlingual links.

The posts in the EFG WikiCorpus are written by more than 2 million different contributors, including bots, but excluding anonymous users. For each language, Table 3 provides the top 2 editors and the top 2 post writers with the number of posts they published (#posts) and the total number of edits they performed on WP until August 2019[28] (#edits). Bots are excluded from these numbers.

	lang	pseudo user	#posts	#edits
#edits top 1	E	Ser Amantio di Nicolao	0	5441797
	F	Polmars	383	1019479
	G	Harry8	11518	566311
#edits top 2	E	BrownHairedGirl	122	2894255
	F	Vlaam	670	909350
	G	Invisigoth67	2598	374693
#posts top 1	E	Will Beback	25078	112162
	F	Jean-Jacques Georges	10119	208779
	G	Kopilot	27126	95046

[26] For the F WikiCorpus, archives are mostly npov pages with 1740 archive pages and 3096 npov pages.

[27] A word is defined as a sequence of alphanumeric characters with the character "_" i.e. using the \w+ perl regular expression.

[28] This number is given by WP and combines the edits on all namespaces i.e. article, talk, user, category, etc. pages: https://en.wikipedia.org/wiki/Wikipedia:List_of_Wikipedians_by_number_of_edits

#posts top 2	E	Jayjg	24191	134742
	F	Racconish	8905	61944
	G	Phi	26348	73544

Table 3: Top 2 editors and top 2 posts in the WP.en, WP.fr and WP.de

Surprisingly, the most prolific post writers are not the ones who edit massively. In other words, people who talk behind the articles are not people who edit articles.

As stated before, multilingual links between *tp*s are made according to the links between article, portal and category pages. Table 4 gives the number of *tp*s for which article counterparts in one or both of the other languages are encoded in the <TEI:relatedItem> elements (see Section 3.2). "# both links" indicates the number of *tp*s for which the article has a counterpart in the two other languages. "# no link" indicates the number of *tp*s for which there is no article counterpart.

	# no link	# en link	# fr link	# de link	# both links
E WikiCorpus	19,396		461,768	417,380	270,717 (13%)
F WikiCorpus	1,751	135,480 (51%)		97,439	87,610 (33%)
G WikiCorpus	2,242	369,582 (52%)	268,029		240,228 (34%)

Table 4: Number of *tp*s according to the number of links to an article counterpart in one or both of the other languages as encoded in the TEI:relatedItem elements.

Not surprisingly, links to English are the most widespread. More than half of WP.fr and WP.de *tp*s have a link to the counterpart article in English. In contrast, only 20% of the WP.en *tp*s have a link to the French or the German counterpart article, and only 13% have both links. Generally, it is quite rare to have no link or only one link to the French or German counterpart article.

The multilingual links to the portals and categories are more difficult to describe from scratch. We could just notice that main portals are rarely indicated and that the top 3 main portals are ‘Geography’, ‘Culture and the arts’ and ‘Society and social sciences’. The category labels are currently too diverse for a comparable overview to be proposed.

4.3- Brief linguistic overview

This section proposes a brief overview of the linguistic content of the EFG WikiCorpus. The main objective is to describe some broad characteristics of the corpus in terms of interactional and linguistic content. A more detailed analysis of the discussions that take place in the EFG WikiCorpus is proposed in Chapter 2.1. This brief overview is based on a look at the 100 most frequent words in each language and in various configurations: everywhere in the tp, in the heading of the threads, in the first and the second posts of the threads and in the beginning and the ending of the posts. The top ranked lemmas differ totally between the three languages (cf. Table 5). In English, the most frequent lemma is "UTC" (standing for Coordinated Universal Time and occurring necessarily in signatures) followed by the lemma "article". In French, it is the negative particle "ne" followed by its natural complement "pas". In German, it is the connective "auch" (*also*) followed by "CE(S)T"

(standing for Central European (Summer) Time, as in English). The first ranked common lemma in all three languages is "article", followed by "have"/"avoir"/"haben".

We may interpret and distinguish three main classes among these most frequent lemmas. The first class includes words referring to the main generic topics discussed between users, i.e., the article page and sections, the sources cited (or not sufficiently cited) in the article following one of the pillar rules of WP, or the discussion itself. These words are very frequent in all configurations and all languages. A second class includes words that describe editing actions. These words occur more in headings than elsewhere. A third class could be identified by grouping words that seem to be used for interacting with the other users, e.g., addressing, greeting, being polite, asking why and (dis)agreeing. This last class includes words that occur particularly frequently in the initial position of first posts.

Table 5 provides the relative frequency and the rank (in brackets) of the most frequent lemmas of the whole EFG WikiCorpus that could be linked to these classes. Some cells are empty when there is no simple term for expressing the class in the top 100 lemmas. This is typically the case for action and interaction classes that occur specifically in special configurations (headings or beginning of first posts) without a very high frequency in the entire EFG WikiCorpus. The last line gives the number of occurrences of negative particles, as they are surprisingly the two most frequent words in the French subcorpus.

	Lemmas in	**Relative freq. per million words (rank)**		
Class	**English, French, German**	**E WikiCorpus occ.**	**F WikiCorpus occ.**	**G WikiCorpus occ.**
Topic	*article (en, fr), Artikel*	5564.0 (2)	6082.3 (4)	5007.4 (3)
	page (en, fr), Seite	2552.2 (6)	1987.7 (15)	767.3 (49)
	section (en, fr), Abschnitt	1588.7 (22)	706.9 (76)	974.8 (37)
	source (en, fr), Quelle	2489.2 (7)	2408.3 (13)	1367.2 (22)
	discussion (en, fr), Diskussion	739.9 (77)	768.2 (66)	771.8 (48)
Action	*change, mettre, machen*	1590.0 (21)	1201.6 (28)	1322.7 (24)
	write, écrire, schreiben	760.0 (72)	803.5 (62)	927.5 (40)
	add, ajouter, hinzufügen	1253.8 (30)	745.6 (72)	
	remove, supprimer, löschen	901.4 (53)	622.1 (99)	
Interaction	*hello, bonjour, hallo*		949.0 (47)	
	please, (s'il vous)plaît, bitten	2067.8 (14)		684.8 (58)
	thx/thank, merci, danke		794.3 (63)	
	agree, accord, doch	699.8 (90)		1198.7 (28)

	why, pourquoi, warum		620.1 (100)	629.3 (70)
neg. part.	*not, ne/pas, nicht/kein*	1099.8 (37)	14074.8 (1) 11489.9 (2)	

Table 5: Rank and relative frequency of selected lemmas among the most frequent lemmas occurring in the EFG WikiCorpus.

As indicated in Section 3.2.3, template processing permits the encoding of emojis and addressing (e.g., reply-to, ping) as special TEI features. Table 6 provides the relative frequency per million words of emoji and addressee templates in the EFG WikiCorpus.

	Relative freq. per million words			
	E WikiCorpus occ.	**F WikiCorpus occ.**	**G WikiCorpus occ.**	**EFG WikiCorpus occ.**
emojis	1.7	279.6	0.6	21.3
addressing	78.1	553.2	71.7	110.6

Table 6: Relative frequency per million words of the emoji and addressing templates in the EFG WikiCorpus

French users are the contributors who seem to use emoji and addressing templates the most, although we are not able to provide any explanation without further investigations. A first requirement will be to complete this inventory with the emojis directly written with combinations of punctuation marks, in addition to the use of templates.

4.4. The EFG WikiDemoCorpus (WDC): A derived subcorpus for more qualitative analyses

To facilitate qualitative analyses, a derived subcorpus was built: the EFG WikiDemoCorpus (WDC). This subcorpus is composed of article and talk pages related to diverse controversial topics that are relevant for all languages and that show significant activity. The controversial nature of the topic is based on the presence of a banner in the WP.en, WP.fr or WP.de *tp*'s header warning that "The subject of this article is controversial and content may be in dispute". Significant activity means that there is at least one archive *tp* and a high number of posts in the current *tp*. In addition, some of these talk pages have already been studied in previous linguistic studies (e.g., Poudat et al. 2016, Poudat 2017).

Table 7 presents the composition of this EFG WDC. We can see that data are more or less balanced among languages depending on the topic (bold font highlights the numbers that are significantly higher for one language).

Portal	Article title	#talk pages	#words in talk pages

	(WP.en, WP.fr, WP.de)						
		E	F	G	E	F	G
EFG WikiDemoCorpus (WDC)		162	33	41	3949521	719334	762049
Politics	European_migrant_crisis, Crise_migratoire_en_Europe, Flüchtlingskrise_in_Europa_ab_2015	2	1	6	35430	9708	**179527**
Biology	Chiropractic, Chiropratique, Chiropraktik	41	3	2	**1709125**	33533	13825
People	Vladimir_Putin Vladimir_Poutine Wladimir_Wladimirowitsch_Putin	16	3	5	**345205**	57531	95709
History	September_11_attacks Attentats_du_11_septembre_2001 Terroranschläge_am_11._September_2001	64	4	6	**1330399**	133667	135662
Society	Psychoanalysis Psychanalyse Psychoanalyse	6	9	8	55655	93328	**137887**
Technology	Genetically_modified_organism Organisme_génétiquement_modifié Gentechnisch_veränderter_Organismus	4	10	2	54954	**343124**	3693
Society	Feminism Féminisme Feminismus	22	2	10	**357281**	41780	164352
Life	The_Legend_of_Zelda[29]	7	1	2	**61472**	6663	31394

Table 7: EFG WikiDemoCorpus (WDC) composition with portal information as indicated in the WP.en talk page

The EFG WDC is searchable online via the Korap application, which allows different query languages: https://korap.ids-mannheim.de/instance/wikidemo. Figure 7 gives an extract of the results (of the 17,504 hits) obtained with the Poliqarp query searching for occurrences of the lemma "do" in the E WDC only (i.e., "corpusSigle eq WDE19").

[29] All languages use the same article title.

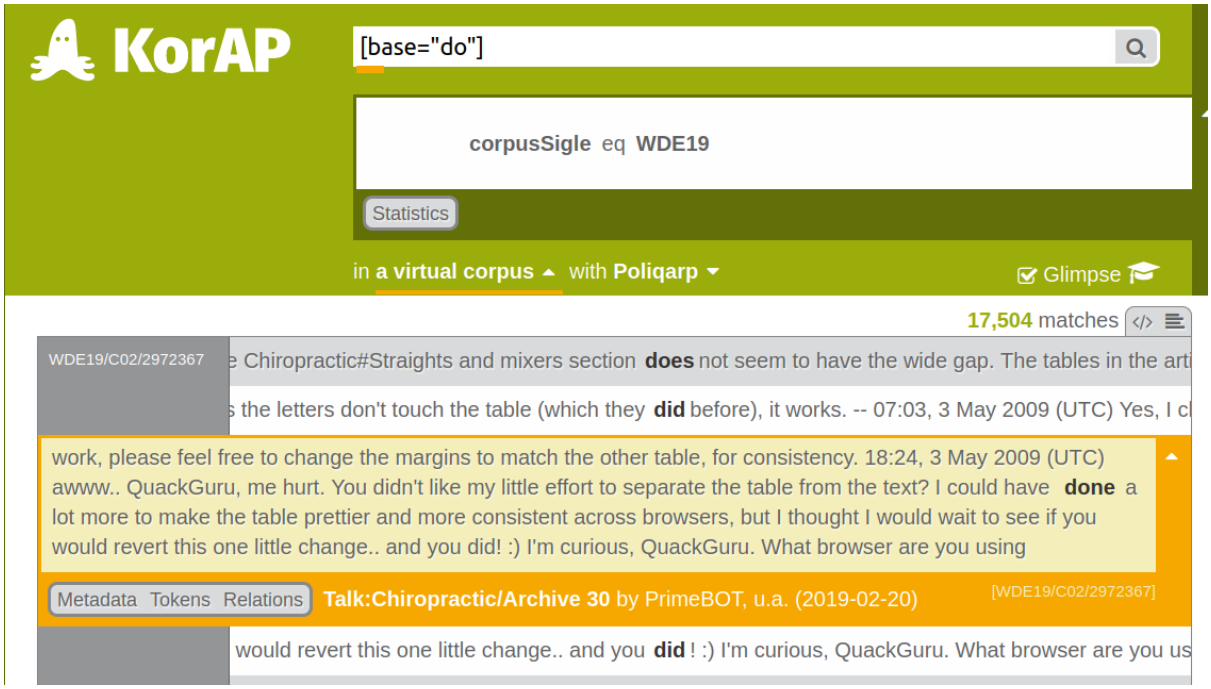


Figure 7: Korap platform for querying the EFG WikiDemoCorpus (WDC)

EFG WDC articles are also available on the Contropedia platform (Borra et al. 2015 and Laniado et al. in this volume), which provides an analysis of all the edits performed in the article pages: https://www.contropedia.net/demo. For example, the WP.en article on Feminism is available at https://www.contropedia.net/demo/index.php?title=Feminism.

5- Conclusion

This chapter presents a method for building a corpus of online discussions related to WP articles in three languages: English, French and German. The resulting corpus is made available for download in the Ortolang platform (Ho-Dac, 2024)[30] and for querying in the Korap platform. The EFG WikiCorpus proposes a comparable corpus with rich metadata and a TEI-CMC core encoding that opens many further application possibilities for linguistic corpus studies. Some of these studies have already been initiated and are explained in the following chapter of this volume. Some of them use the entire EFG WikiCorpus (Tanguy et al., Chap. XX in this volume), the derived EFG WDC (Gredel, Chap. XX in this volume) or a monolingual subpart of it: the German part in Herzberg and Lüngen (Chap. XX in this volume) and the French part in Carbou et al. (Chap. XX in this volume).

These studies explore various aspects of the talk pages: the social interactions of which these talk pages are the trace; the way people discuss when working together; the linguistic expression of specific speech acts (e.g., conflict, addressing, agreement); the encyclopedic textual genre and the ideological regime this genre involves. All these studies accept the challenge of mining and characterizing a new kind of language resource that is part of computer-mediated communications (CMCs). CMCs are challenging both methodologically and theoretically, cf. the International Conference Series on CMC and Social Media Corpora

[30] https://hdl.handle.net/11403/efg-wikicorpus

(https://cmc-corpora.org). In this chapter, we try to bring some order into the WP world and to render the EFG Wiki(Demo)Corpus attractive for linguistics, conversation analysis, communication sciences and human sciences in general.

References

Baldwin, Timothy, Cook, Paul, Lui, Marco, MacKinlay, Andrew & Wang, Li. 2013. How noisy social media text, how different social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing,* Ruslan Mitkov & Jong C. Park (eds), 356–364. Nagoya, Japan.

Beißwenger, Michael & Lüngen, Harald. 2020. CMC-core: A schema for the representation of CMC corpora in TEI. *Corpus 20*. 〈http://journals.openedition.org/corpus/4553〉

Beißwenger, Michael, Wigham, Ciara, Etienne, Carole, Grumt Suárez, Holger, Herzberg, Laura, Fišer, Darja, Hinrichs, Erhard, Horsmann, Tobias, Karlova-Bourbonus, Natali, Lemnitzer, Lothar, Longhi, Julien, Lüngen, Harald, Ho-Dac, Lydia-Mai, Parisse, Christophe, Poudat, Céline, Schmidt, Thomas, Stemle, Egon, Storrer, Angelika & Zesch, Torsten. 2017. Connecting resources: Which issues have to be solved to integrate CMC corpora from heterogeneous sources and for different languages? In *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (Cmccorpora17)*, Egon W. Stemle & Ciara Wigham (eds) 52–55. Bolzano, Italy.

Borra, Erik, Weltevrede Esther, Ciuccarelli, Paolo, Kaltenbrunner, Andreas, Laniado, David, Magni, Giovanni, Mauri, Michele, Rogers, Richard & Venturini, Tommaso. 2015. Societal controversies in wikipedia articles. In *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 193–196. New York, NY: ACM.

Chang, Jonathan P., Chiam, Caleb, Fu, Liye, Wang, Andrew Z., Zhang, Justine & Danescu-Niculescu-Mizil, Cristian. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt & Stefan Ultes (eds), 57–60. [System demo]. Stroudsburg PA: ACL.

Chang, Jonathan P. & Danescu-Niculescu-Mizil, Cristian. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (XXth EMNLP)*. Stroudsburg PA: ACL.

Elia, Antonella. 2009. Quantitative data and graphics on lexical specificity and index readability: The case of wikipedia. *Revista Electrónica de Lingüística Aplicada 8*: 248–271.

Ferschke, Oliver, Gurevych, Iryna & Chebotar, Yevgen. 2012. Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 777–786. Stroudsburg PA: ACL.

Ho-Dac, Lydia-Mai. 2024. EFG WikiCorpus — discussions in Wikipedia's backstage (English, French, German) [Corpus]. *ORTOLANG* (Open Resources and TOols for LANGuage) — www.ortolang.fr, https://hdl.handle.net/11403/efg-wikicorpus

Ho-Dac, Lydia-Mai & Laippala Veronika. 2017. Le corpus WikiDisc: Ressource pour la caractérisation des discussions en ligne. In *Corpus de communication médiée par les réseaux: Construction, structuration, analyse*. Ciara R. Wigham & Gudrun Ledegen (eds), 107–124. Paris: l'Harmattan.

Ho-Dac, Lydia-Mai, Laippala, Veronika, Poudat, Céline & Tanguy, Ludovic. 2017. Exploring Wikipedia talk pages for conflict detection. In *Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World*, Darja Fišer & Michael Beißwenger (eds), 146–168. Ljubljana: Ljubljana University Press, Faculty of Arts.

Huta, YiqingDanescu-Niculescu-Mizil, Cristian, Taraborelli, Dario, Thain, Nithum, Sorensen, Jeffery & Dixon, Lucas. 2018. WikiConv: A corpus of the complete conversational history of a large online collaborative community. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, 2818–2823. Stroudsburg PA: ACL.

Konieczny, Piotr. 2010. Adhocratic governance in the internet age: A case of Wikipedia. *Journal of Information Technology & Politics 7*(4): 263–283.

Laniado, David, Tasso, Riccardo, Volkovich, Yana & Kaltenbrunner, Andreas. 2011. When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 11)*, Barcelona, 17–21 July.

Langlais, Pierre-Carl. 2014. La négociation contre la démocratie : le cas Wikipedia. *Négociations 1*: 21–34.

Lehmann, Jens, Isele, Robert, Jakob, Max, Jentzsch, Anja, Kontokostas, Dimitri, Mendes, Pablo N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S. & Bizer, C. 2015. Dbpedia — A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web,* 6(2), 167–195.

Lydia-Mai Ho-Dac Lüngen, Harald & Herzberg, Laura. 2019. Types and annotation of reply relations in computer-mediated communication. *European Journal of Applied Linguistics 7*(2): 305–331.

Margaretha, Eliza & Lüngen, Harald. 2014. Building linguistic corpora from Wikipedia articles and discussions. *Journal for Language Technology and Computational Linguistics 29*(2): 59–82.

Medelyan, Olena, Milne, David, Legg, Catherine & Witten, Ian H. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Interactions 67*(9): 716–754.

Mintzberg, Henry. 1979. *The Structuring of Organizations*. Englewood Cliffs NJ: Prentice-Hall.

Mitrevski, Blagoj, Piccardi, Tiziano, & West, Robert. 2020. WikiHist.html: English Wikipedia's full revision history in HTML Format. *Proceedings of the International AAAI Conference on Web and Social Media* 14: 878–884.

Myers, Greg. 2010. The Discourse of Blogs and Wikis. London: Continuum.

Poudat, Céline, Grabar, Natalia, Paloque-Bergès, Camille, Chanier, Thierry & Jin, Kun. 2017. Wikiconflits: Un corpus de discussions éditoriales conflictuelles du Wikipédia francophone. In *Corpus de communication médiée par les réseaux: Construction, structuration, analyse,* Ciara R. Wigham & Gudrun Ledegen (eds). Paris: l'Harmattan.

Poudat, Céline, Vanni, Laurent, & Grabar, Natalia. 2016. How to explore conflicts in French wikipedia talk pages? In *Statistics Analysis of Textual Data*, Nice, France, June, 645–656. 〈https://hal.science/hal-01359416/document〉 (1 June 2024).

Potthast, Martin, Stein, Benno, Gerling, Robert. 2008. Automatic Vandalism Detection in Wikipedia. In Advances in Information Retrieval. ECIR 2008. *Lecture Notes in Computer Science*, Vol. 4956, Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven & Ryen W. White (eds), 663–668. Springer, Berlin, Heidelberg.

Walton, Aengus. 2009. *A Statistical Analysis of Stylistics and Homogeneity in the English Wikipedia*. PhD dissertation, Trinity College Dublin.

Wulczyn, Ellery, Thain, Nithum and Dixon, Lucas. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web,* 1391–1399. International World Wide Web Conferences Steering Committee.

Zesch, Torsten, Müller, Christof & Gurevych, Iryna. 2008. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International*

Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. Paris: European Language Resources Association (ELRA).

Zhang, Justine, Chang, Jonathan P., Danescu-Niculescu-Mizil, Cristian, Dixon, Lucas, Hua, Yiqing, Thain, Nithum & Taraborelli, Dario. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Vol. 1: Long Papers,* Iryna Gurevych & Yusuke Miyao (eds), 1350–1361. Stroudsburg PA: ACL.